Spring 3-12-2021

Identifying, Analyzing, and Using Discriminatory Variables for Classification of Neutrino Signal and Background Noise in Multivariate Analysis in the Askaryan Radio Array Experiment

Jesse Osborn
University of Nebraska - Lincoln

Identifying, Analyzing, and Using Discriminatory Variables for Classification of Neutrino Signal and Background Noise in Multivariate Analysis in the Askaryan Radio Array Experiment

An Undergraduate Honors Thesis
Submitted in Partial fulfillment of
University Honors Program Requirements
University of Nebraska-Lincoln

By
Jesse Osborn, BS
Physics and Mathematics
College of Arts and Sciences

Faculty Mentor: Dr. Ilya Kravchenko, Department of Physics and Astronomy

March 12, 2021

Appreciation

I would like to thank my advisor, Dr. Kravchenko, for being patient with me as I have developed my research skills over the past nearly three years under his guidance and for working with me every step of the way on this project. I would also like to thank previous UNL student Andrew Schultz for his help developing software for this project, as well as other members of the Askaryan Radio Array Experiment and the UNL Physics Department's High Energy Group for their valuable feedback on this work as it was in development. Lastly, I would like to thank UNL's Undergraduate Creative Activities And Research Experience (UCARE) Program for funding the majority of my work for Dr. Kravchenko.

Abstract

The Askaryan Radio Array Experiment, located near the South Pole, works to pinpoint specific instances of neutrinos from outside the solar system interacting with nucleons inside the Antarctic ice, emitting radio waves. I have taken data from the ARA stations which is presumed to be background noise and compared it to simulated data meant to look like a neutrino signal. I developed a suite of variables for discrimination between the two data sets, using a computer algorithm to generate a single output variable which can be used to distinguish noise events from signal events. I maximized this discrimination process for simulated neutrino energies between 10^{17} and 10^{19} electron volts.

Key Words: Physics, neutrino, high energy, machine learning

1: Introduction

The Askaryan Radio Array (ARA) is a large-scale detector at the South Pole concerned with the measurement of ultra-high energy (UHE) neutrinos and their flux here on Earth through the Askaryan Effect. The Askaryan Effect is a process by which neutrinos generate a shower of particles while passing through the Antarctic ice which emits radio wave radiation, first experimentally observed in 2000.

The ARA detector currently consists of 5 stations (or modules), each consisting of 16 radio antennas (or channels) submerged roughly 200 meters deep in the Antarctic ice. The first 3 stations were built in 2012, and the most recent 2 were installed at the end of 2017 using National Science Foundation (NSF) funding.

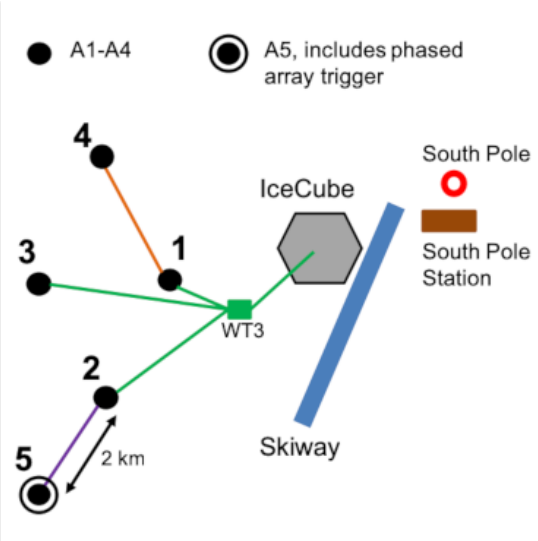

Figure 1: Current arrangement of ARA stations near the South Pole

ARA stations listen for a weak radio frequency (RF) pulse, a sign of a neutrino interacting with an ice molecule. A station is triggered and records an event when enough power is observed on three or more of its channels in a short time interval, indicating some substantial signal was received by the station. Since the neutrino signal is weak, the power threshold for each antenna is low enough that many thermal noise and anthropogenic events can also be recorded while looking for neutrino signals. Currently, ARA has three filters to eliminate these non-neutrino events: The Hit Filter, which requires that at least 4 of a station's 16 channels have a hit (recording power above their thresholds), the Time Sequence Filter, which requires that the hit antennas observe hit time differences consistent with a plane wave approaching the station, and the Arrival Angle Filter which excludes hits from above the surface of the ice. Although these filters exclude many non-neutrino events, they do not exclude all of them. Therefore, it is crucial to find new ways in which to filter out noise events while retaining legitimate neutrino signal events.

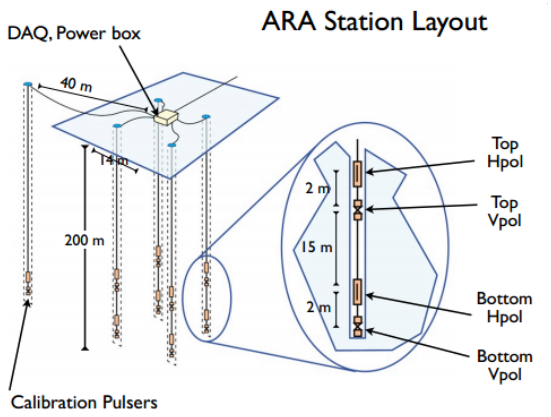

Figure 2: Design of ARA stations and placement of radio wave antennas in the Antarctic ice.

Studying these UHE neutrinos allows testing of various theories that cannot be tested in any other way. Because of their extremely small mass and neutral electric charge, neutrinos rarely interact with other matter as they travel through space, so they provide a means to study processes occurring far beyond our solar system that are otherwise inaccessible. This research focused on developing an additional noise filter to improve the neutrino signal detection capabilities of ARA, making use of multivariate analysis techniques not previously incorporated in ARA's other noise filters.

2: Excursion Finder

For these filter studies, the unblinded 10 percent of data collected from ARA station 2 in 2014 was passed through the three current noise filters, and 2728 events survive. The expectation from existing theory is that the vast majority of these events are still noise, as real neutrino signals are only expected to be recorded a handful of times per year. As such, these 2728 events were treated as a sample of background noise. An example of such an event can be seen in Fig. 3.

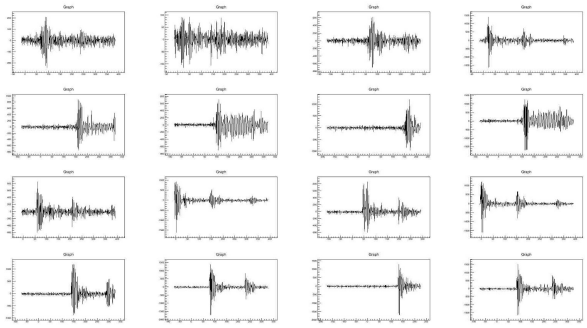

Figure 3: Example of an event recorded by the stations that passed all existing filters. Each of the 16 plots is a voltage vs time graph.

The background noise was compared to 20,000 simulated 10^{19} eV neutrino events with neutrinos generated using AraSim, a software package developed by ARA. These 20,000 events were treated as a sample of neutrino signal. An example of such an event can be seen in Fig. 4.

The first necessary element to this study was to find a way to distinguish something interesting happening in a particular channel of an event from nothing at all. We want to have a program that can effectively identify the interesting parts of events for us so we don't have to comb through thousands of events and sort them by hand. Hence, I developed an "excursion" finder, where an excursion is just any amount of voltage recorded by an antenna that is significantly above the typical noise recorded by that channel. Certainly, if anything interesting did happen in that event, at least one of the 16 channels will meet this requirement.

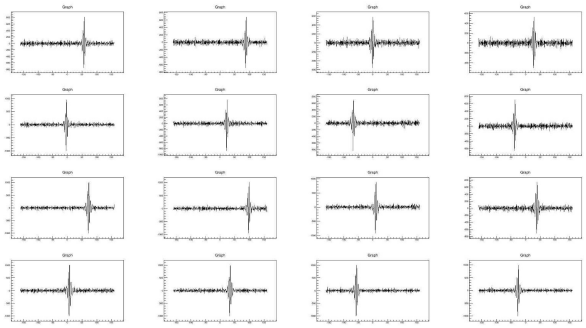

Figure 4: Example of a simulated event. Each of the 16 plots is a voltage vs time graph.

The specific programming of this excursion finder was as follows. First, because the nature of the data I am observing is voltage vs time plots and voltage can easily have significant switching between positive and negative values from data point to data point, I use an envelope to smooth out the data. The envelope of choice for my excursion finder was the Hilbert Envelope, which transforms each data point of the raw signal as such

$$e_i = \sqrt{v_i^2 + h_i^2}, \tag{1}$$

where e_i, v_i, and h_i are the ith data points of the envelope, raw signal, and Hilbert Transform respectively.

The procedure for using the envelope to define an excursion is as follows:

1. Break the envelope into four time quadrants

2. Threshold level for each waveform is equal to 4.0 times the average mean voltage value of the two smallest envelope quadrants (noise level)

3. An excursion is any number of sequential points above this threshold level

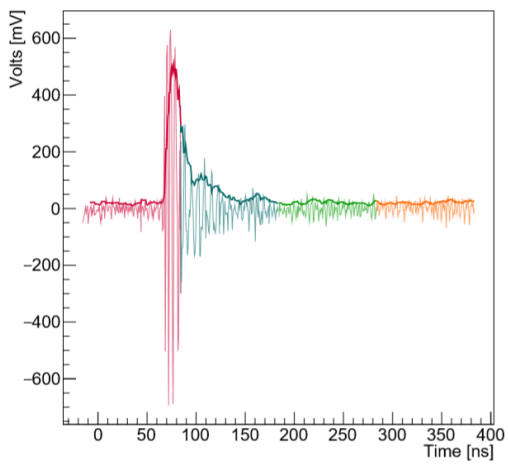

Figure 5: Example of a waveform with an envelope calculated to build a threshold being divided into four quadrants. In this case, the green and orange quadrants would be used to define the noise level.

3: Multivariate Analysis Parameters

Using the excursion finder as the primary point of analysis, we set out to develop a collection of calculable parameters that were not typically the same between noise and signal events. Initially, by visual inspection, we compared the signal and noise events by eye to see what features seemed to be characteristic of noise events and not signal events. Things like how many excursions were recorded for each event or for each of the 16 channels making up the event, how long these excursions were, how early or late in the channels' total recording they were, etc. Through trail and error looking at the distributions of variables between signal and noise events, we discarded some of our initial ideas. The final list of the parameters used for our multivariate analysis is given below. Note: variables that deal with a single excursion deal with the first excursion that occurs in a given channel (all references to an excursion are to the first excursion).

1. totMax - The longest excursion in the event (time over threshold)
2. exPerEvent - The total number of excursions per event
3. nExcUntilEnd - The number of excursions that last until the end of a channel's recording
4. nExcChanMax - The largest number of excursions in a single channel
5. firstDeltaTMax - The excursion farthest from the start of its channel's recording in the event
6. firstDeltaTMin - The excursion closest to the start of its channel's recording in the event
7. lastDeltaTMax - The excursion that is farthest away from the end of its channel's recording in the event
8. lastDeltaTMin - The excursion that is closest to the end of its channel's recording in the event
9. trailDeltaTMax - The excursion with the longest time between the end of the excursion above a threshold equal to 6 times the noise level and the end of the excursion above the normal threshold level (4 times the noise level) in the event
10. trailDeltaTMin - The excursion with the shortest time between the end of the excursion above a threshold equal to 6 times the noise level and the end of the excursion above the normal threshold level (4 times the noise level) in the event

Note that trailDeltaTMax and trailDeltaTMin may not always be well defined for every event, since the excursion finder only requires that events have an excursion at the normal threshold level (4 times the noise level). These excursions may not rise to a threshold level set at 6 times the noise level, so these two variables are undefined more often than the others. In the event that these variables, or any others, are undefined, they are given default values that would not be seen in actual events (either extremely large positive or extremely large negative values), and they are barred from use in the multivariate analysis. The distributions for these variables between the signal and noise samples can be seen in Fig. 6.

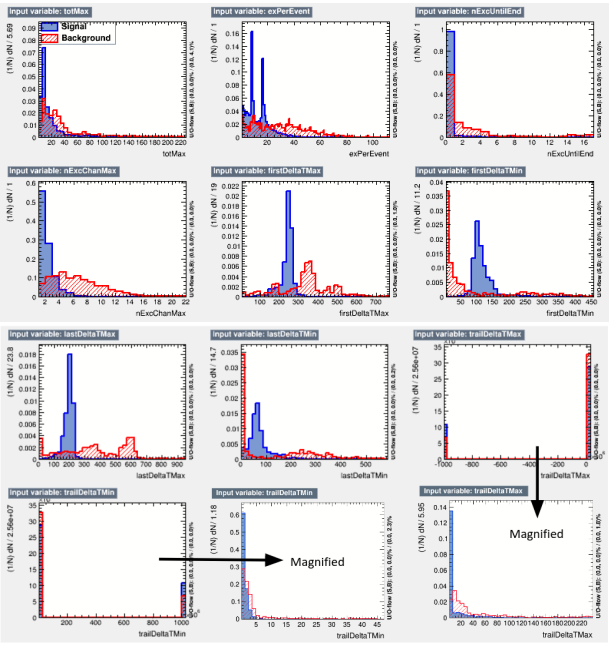

Figure 6: Distribution of multivariate analysis parameter values between signal (blue) and background (red) samples. The variables that the distributions belong to run from #1 (top left) to #10 (bottom right) for the variables as given in the list on the previous page.

The statistics of these variables (considering signal and background distributions together) can be seen in Table. 1. All variables are either counts of excursions or measurements of time in nanoseconds.

Table 1: Input Variable Statistics

Input Variable	Mean	RMS	Min	Max
totMax	18.32	27.415	5.4e-03	278.9
exPerEvent	14.04	10.727	1.000	120.0
nExcUntilEnd	0.2551	1.4531	0.000	16.00
nExcChanMax	2.176	2.0473	1.000	23.00
firstDeltaTMax	248.6	59.740	6.068	813.9
firstDeltaTMin	110.4	44.500	4.692	514.7
lastDeltaTMax	210.1	88.845	0.000	926.2
lastDeltaTMin	73.09	62.667	0.000	749.5
trailDeltaTMax	-2.5e+08	4.4e+08	-1.0e+09	274.0
trailDeltaTMin	2.5e+08	4.4e+08	0.14759	1.0e+09

4: Multivariate Analysis Execution

Now that we have developed some parameters that show some promise (by eye) for discriminating between signal and background noise, we can use a computer algorithm to perform our multivariate analysis. The software of choice for this study was the Toolkit for Multivariate Analysis (TMVA), a ROOT-integrated environment for the processing, parallel evaluation, and application of multiple classifying variables.

TMVA is home to various different multivariate analysis (MVA) techniques for dealing with classifying variables. TMVA accepts and analyses classifying variables for defined signal and background training events according to the selected MVA technique. TMVA then determines the most effective discrimination methods by applying weighting to each of the input parameters and applies these discrimination methods to defined signal and background testing events to evaluate the discrimination efficiency. TMVA uses half of the background noise sample (1364 events) and half of the simulated neutrino signal sample (10,000 events) for training events and the other halves for testing events.

The particular MVA technique used for this study was the Boosted Decision Trees (BDT) method, in which TMVA applies numerous cuts to each variable's training data, splitting the variable into various small nodes that are eventually classified as either purely signal or purely background. This is the decision trees aspect of BDT. TMVA then re-weights the training data and performs the cuts again, eventually taking a weighted average of all the iterations and merging them into one set of cuts. This is the boosted aspect of BDT. A visual representation of the BDT method can be seen in Fig. 7.

In essence, TMVA takes all ten of our input variables and decides (using the BDT method) the most optimal way to combine that information and generate a single output variable on a per-event basis, which we can then perform a simple cut on to remove background noise. The output variable generated for our

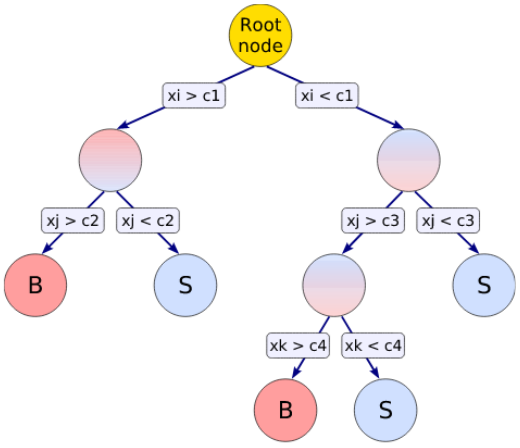

Figure 7: Representation of the decision tree process used by the TMVA BDT algorithm on each of the variables to segment the distributions into small "purely signal" and "purely background" nodes.

ten input variables on our signal and noise samples can be seen in Fig. 8. Note that TMVA, numerically, tries to give all background noise events an output variable approaching -1 and all simulated neutrino signal events an output variable approaching +1.

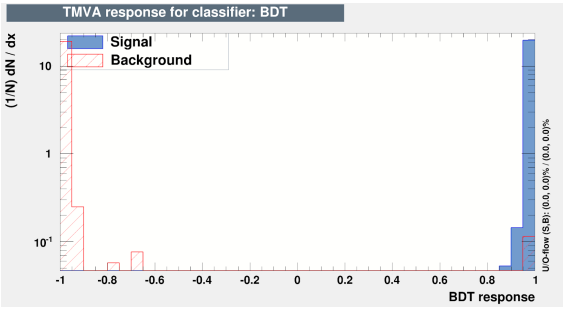

Figure 8: Output variable generated by TMVA using the Boosted Decision Tress method on the ten input variable distributions seen in Fig. 6. The y axis has been given log scaling.

TMVA generates a lot of statistics about our input and output variables that can be used to improve upon the process. One such statistic is the efficiency with which we can cut background noise events versus how many signal events would be lost (performing a linear cut at different points in Fig. 8. This can be seen in Fig. 9.

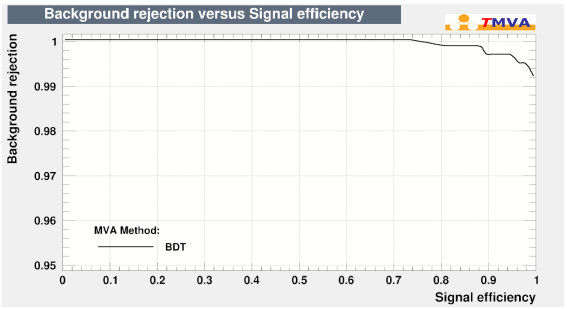

Figure 9: Receiver Operating Characteristic (ROC) curve generated by TMVA which details the background rejection ability of the output TMVA variable seen in Fig. 8.

From Fig. 9, we see that we can achieve a 99.75% reduction in background signal while maintaining 95.1% signal efficiency for our testing events.

Another statistic TMVA gives us is the evaluated importance of each of the input variables in getting this performance in the output variable, which can be seen in Table. 2.

Table 2: Ranked Input Variable Importance

Rank	Input Variable	Importance
1	lastDeltaTMax	1.857e-01
2	exPerEvent	1.656e-01
3	firstDeltaTMax	1.509e-01
4	firstDeltaTMin	1.464e-01
5	lastDeltaTMin	9.777e-02
6	nExcChanMax	7.878e-02
7	trailDeltaTMax	6.414e-02
8	totMax	5.737e-02
9	trailDeltaTMin	2.826e-02
10	nExcUntilEnd	2.515e-02

One last statistic that TMVA gives us is correlation matrices for our variables for both the background and signal samples. These matrices tell us how much of a relationship there is between our different variables values from one event to another. Both of these correlation matrices can be seen in Fig. 10.

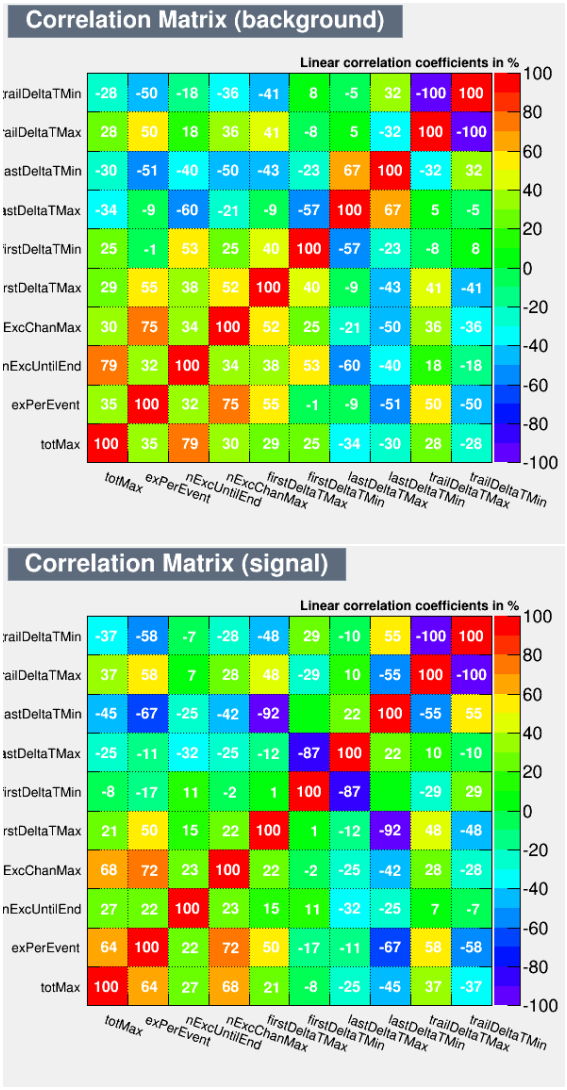

Figure 10: Correlation matrices for the input variables to TMVA for the background sample (top) and signal sample (bottom) pictured in Fig. 6.

The results of Table. 2 and Fig. 10 can be used to optimize the performance of our filter. If we have a variable that is barely used by TMVA for discrimination or if two variables give TMVA the exact same information for the purposes of discrimination, then it's not worthwhile to invest computation in calculating both variables for the millions of events that a noise filter would be used on.

For the purposes of this research, no further modifications were made to the TMVA process, although future research may find it worthwhile to either add to or subtract from the input parameters discussed here based on these statistics.

5: Energy Dependence

In the multivariate analysis up to this point, we had been working with one noise sample (2728 events taken from real data) and one signal sample (20,000 simulated events). These simulated events were all generated assuming a neutrino energy of 10^{19} eV. For this simulated sample, the multivariate analysis techniques discussed above seemed to be capable of performing a significant reduction in noise events.

But, the question of whether the multivariate techniques are only effective at preserving signals of 10^{19} eV remains. In order to study this question, we generated additional simulated neutrino signal samples assuming neutrino energies of 10^{18} eV and 10^{17} eV and performed the exact same TMVA analysis that we did on our initial signal sample. We varied both the energy of simulated signal that TMVA trained on (which creates the weighting of the input variables that is used for testing) and the energy of simulated signal that TMVA actually performed on. For example, if we train TMVA on 10^{17} eV signal, do we still get the same noise discrimination effectiveness if we then use that trained TMVA on a 10^{19} eV signal sample?

The output variable for each of the three different signal samples was calculated using the input variable weighting from TMVA training on the 10^{17} eV simulated signal sample, the 10^{18} eV simulated sig-

nal sample, and the 10^{19} eV simulated signal sample, for a total of nine different output variable calculations (plus an additional 3 for the background sample, which was the same for all TMVA training). All of these can be seen in the three plots in Fig. 11.

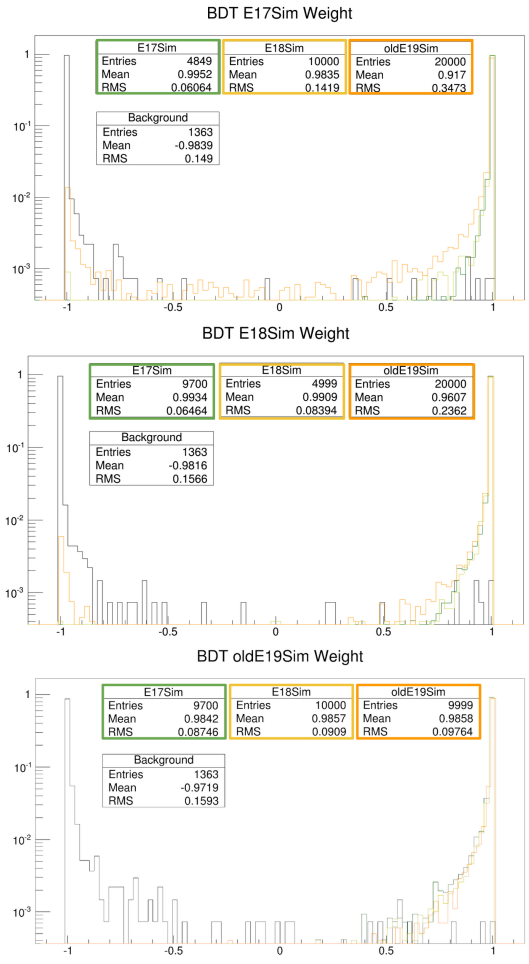

Figure 11: Comparison of output variables computed for simulated signal samples with energies of 10^{17} eV, 10^{18} eV, and 10^{19} eV, where the input variable weighting was taken from the TMVA training with the 10^{17} eV simulated sample (top), the 10^{18} eV simulated sample (middle), and the 10^{19} eV simulated sample (bottom). The y axes have been given log scaling.

The results of the three plots in Fig. 11 are summarized in Tables. 3, 4, 5, and 6.

Table 3: Energy Study Summary - 10^{17} eV Signal

Training Weight	BDT Variable Mean
10^{17} eV	0.9952
10^{18} eV	0.9934
10^{19} eV	0.9842

Table 4: Energy Study Summary - 10^{18} eV Signal

Training Weight	BDT Variable Mean
10^{17} eV	0.9835
10^{18} eV	0.9909
10^{19} eV	0.9857

Table 5: Energy Study Summary - 10^{19} eV Signal

Training Weight	BDT Variable Mean
10^{17} eV	0.917
10^{18} eV	0.9607
10^{19} eV	0.9858

Table 6: Energy Study Summary - Background Noise

Training Weight	BDT Variable Mean
10^{17} eV	-0.9839
10^{18} eV	-0.9816
10^{19} eV	-0.9719

From these energy comparisons, we can see that there is indeed a marginal difference in the BDT output variable calculated for a signal sample based on the training weight used by TMVA. Naturally, a 10^{17} eV signal sample will receive the best (closest to 1 / farthest from background) BDT variable distribution if the TMVA analysis performed on it was trained using a 10^{17} eV signal sample, and the same holds true for 10^{18} eV and 10^{19} eV signal samples as well. We also note that a 10^{19} eV signal sample will face a substantially larger difference in its BDT output variable if

the TMVA training was done not on its native energy level than 10^{17} eV and 10^{18} eV signal samples will. Further research should investigate whether this trend continues for 10^{20} eV signal samples, in which case it may be possible that these TMVA techniques do become significantly less efficient for noise discrimination at higher energies than we considered in this research.

In summary, if we trained our TMVA filter methodology on signal samples that were of a different energy than that of actual neutrinos, our filter would perform worse for discriminating between neutrinos and noise, although the effect is marginal if the neutrinos have an energy of 10^{17} eV or 10^{18} eV, and only starts to become somewhat significant for energies of 10^{19} eV. Also important to note is that the output BDT variable for the background noise can also be mildly effected by training on different energies of signal samples, which should continue to be observed if further energy dependence studies are conducted.

6: Conclusions

ARA currently has three noise filters: the Hit Filter, Time Sequence Filter, and the Arrival Angle Filter, all of which have been tuned to reject known noise events that are recorded by the stations. ARA station 2 recorded 10,406,300 events in 2014, of which only a handful are expected to result from actual neutrinos. The existing filers reduced this number down to 2728 events, a 99.97% reduction of the number of events. While these filters have eliminated a vast amount of noisy data, there is still too much remaining to reasonably go through by hand for the entire experiment.

As such, this research focused on finding new ways to eliminate the remaining noisy data, making use of a multivariate analysis algorithm (TMVA) and a suite of parameters that was developed by a visual comparison of known noise and neutrino signal events. Through trial and error, we were able to arrive at a collection of parameters that, when making use of a particular multivariate technique known as the Boosted Decision Trees method, was able to generate a single output variable. This output variable, when run on the testing samples, allowed for a 99.75% reduction in background signal while maintaining 95.1% signal efficiency.

The energy dependence of these techniques was also investigated for simulated neutrinos ranging between 10^{17} and 10^{19} electron volts. The results of this investigation lead us to conclude that there is a distinct dependency of the filters effectiveness on if it was initially trained on simulated neutrinos of a different energy than actual neutrinos. However, the effect was not enough to discredit the noise discrimination potential of a filter based on these techniques.

While the exact specifications for filtering noise that were examined in this research may not be used as a filter for ARA, this research does show that there is substantial potential in using multivariate analysis techniques in order to discriminate between neutrino signal and background noise. Further investigation and tuning of these techniques and variables may lead to the development of a new filter for the ARA data that can either work in conjunction with the existing three, or perhaps replace them, in order to further reduce the number of noise events. This process will continue until no noise events remain in the data.